

# Robust sparse canonical correlation analysis

Wilms I, Croux C.



# Robust Sparse Canonical Correlation Analysis

Ines Wilms and Christophe Croux

**Abstract:** Canonical correlation analysis (CCA) is a multivariate statistical method which describes the associations between two sets of variables. The objective is to find linear combinations of the variables in each data set having maximal correlation. This paper discusses a method for Robust Sparse CCA. Sparse estimation produces canonical vectors with some of their elements estimated as exactly zero. As such, their interpretability is improved. We also robustify the method such that it can cope with outliers in the data. To estimate the canonical vectors, we convert the CCA problem into an alternating regression framework, and use the sparse Least Trimmed Squares estimator. We illustrate the good performance of the Robust Sparse CCA method in several simulation studies and two real data examples.

**Keywords:** Canonical correlation analysis, penalized regression, robust regression, sparse Least Trimmed Squares

## 1 Introduction

Canonical correlation analysis (CCA), introduced by Hotelling (1936), identifies and quantifies the associations between two sets of variables. CCA searches for linear combinations, called *canonical variates*, of each of the two sets of variables having maximal correlation. The coefficients of these linear combinations are called the *canonical vectors*. The correlations between the canonical variates are called the *canonical correlations*. For more information on canonical correlations analysis, see e.g. Johnson and Wichern (1998, Chapter 10).

*Sparse* canonical vectors are canonical vectors with some of their elements estimated as exactly zero. The canonical variates then only depend on a subset of the variables, those corresponding to the non-zero elements of the estimated canonical vectors. Hence, the canonical variates are easier to interpret, in particular for high-dimensional data sets. Examples of CCA for high-dimensional data sets can be found in, for example, genetics (Gonzalez et al., 2008; Prabhakar and Fridley, 2012; Cruz-Cano and Lee, 2014) and machine learning (Sun et al., 2011).

Different approaches for sparse CCA have been proposed in the literature. Parkhomenko et al. (2009) use a sparse singular value decomposition to derive sparse singular vectors. Witten et al. (2009) develop a

penalized matrix decomposition, and show how to apply it for sparse CCA. Waaijenborg et al. (2008), Lykou and Whittaker (2010), and Wilms and Croux (2013) convert the CCA problem into a penalized regression framework to produce sparse canonical vectors. All these methods are not robust to outliers. A common problem in multivariate data sets, however, is the frequent occurrence of outliers. Therefore, the possible presence of outliers should be taken into account.

Several *robust CCA* methods have been introduced in the literature. Kärner (1991) considers robust CCA using an M-estimator of multivariate location and scatter; Dehon and Croux (2002) use the minimum covariance determinant (MCD, Rousseeuw, 1985) estimator. Asymptotic properties for CCA based on robust estimators of the covariance matrix are discussed in Taskinen et al. (2006). Filzmoser et al. (2000) use a robust alternating regression approach, following the approach of Wold (1968), to obtain the first canonical variates. Branco et al. (2005) extend the method of Filzmoser et al. (2000) to obtain all canonical variates. CCA can also be considered as a prediction problem, where the canonical variates obtained from the first data set serve as optimal predictors for the canonical variates of the second data set, and vice versa. As such, Adrover and Donato (2015) use a robust M-scale to evaluate the prediction quality, whereas the approach of Kudraszow and Maronna (2011) is based on a robust estimator for the multivariate linear model. None of these methods, however, are sparse.

This paper proposes a CCA method that is sparse and robust at the same time. As such, we deal with two important topics in applied statistics: sparse model estimation and the presence of outliers in the data. We consider CCA from a predictive point of view. We use an alternating robust, sparse regression framework to sequentially obtain the canonical variates. We obtain sparse canonical vectors that are resistant to outlying observations by using the sparse Least Trimmed Squares (sparse LTS) estimator of Alfons et al. (2013). We show that the Robust Sparse CCA method has a clear advantage over standard CCA and Sparse CCA by means of a simulation study. The advantage of Robust Sparse CCA over Robust CCA is twofold. First, Robust Sparse CCA provides well interpretable canonical vectors since some of the elements of the canonical vectors are estimated as exactly zero. Secondly, when the number of variables becomes large compared to the sample size, the estimation accuracy of Robust Sparse CCA will be better. Moreover, when the largest number of variables of both data sets exceeds the sample size, Robust Sparse CCA can still be performed.

The remainder of this article is organized as follows. Section 2 considers the CCA problem from a predictive point of view. Section 3 discusses the Robust Sparse CCA algorithm. Section 4 presents simulation results where we compare Robust Sparse CCA to standard CCA, Robust CCA and Sparse CCA. Two real data examples are discussed in Section 4. Section 5 concludes.

## 2 CCA based on alternating regressions

We consider the CCA problem from a predictive point of view, as proposed by Brillinger (1975) and Izenman (1975). Given a sample of  $n$  observations  $\mathbf{x}_i \in \mathbb{R}^p$  and  $\mathbf{y}_i \in \mathbb{R}^q$  ( $i = 1, \dots, n$ ). The two data matrices are denoted as  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$  and  $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_n]^T$ . The estimated canonical vectors are collected in the columns of the matrices  $\hat{\mathbf{A}} \in \mathbb{R}^{p \times r}$  and  $\hat{\mathbf{B}} \in \mathbb{R}^{q \times r}$ . Here  $r$  is the number of canonical vectors. The columns of the matrices  $\mathbf{X}\hat{\mathbf{A}}$  and  $\mathbf{Y}\hat{\mathbf{B}}$  contain the estimates of the realizations of the canonical variates, and we denote their  $j^{th}$  column by  $\hat{\mathbf{u}}_j$  and  $\hat{\mathbf{v}}_j$ , for  $1 \leq j \leq r$ . The objective function defining the canonical vector estimates is

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \underset{(\mathbf{A}, \mathbf{B})}{\operatorname{argmin}} \sum_{i=1}^n \|\mathbf{A}^T \mathbf{x}_i - \mathbf{B}^T \mathbf{y}_i\|^2. \quad (1)$$

The objective function in (1) is minimized under the restriction that each canonical variate  $\hat{\mathbf{u}}_j$  is uncorrelated with the lower order canonical variates  $\hat{\mathbf{u}}_k$ , with  $1 \leq k < j \leq r$ . Similarly for the canonical vectors within the second set of variables. For identification purpose, a normalization condition requiring the canonical vectors to have unit norm is added. Typically, the canonical vectors are obtained by an eigenvalue analysis of a certain matrix involving the inverses of sample covariance matrices. But if  $n < \max(q, p)$ , these inverses do not exist.

We estimate the canonical vectors with an alternating regression procedure using the Least Squares estimator (Wold, 1968). Indeed, if the matrix  $\mathbf{A}$  in (1) is kept fixed, the matrix  $\mathbf{B}$  can be obtained from a Least Squares regression of the canonical variates on  $\mathbf{y}$  (and vice versa for estimating  $\mathbf{A}$  keeping  $\mathbf{B}$  fixed). The Least Squares estimator, however, is not sparse, nor robust to outliers. Therefore, we replace it by the sparse Least Trimmed Squares (sparse LTS) estimator (Alfons et al., 2013). The sparse LTS estimator is a sparse version of the Least Trimmed Squares estimator. Sparse model estimates are obtained by adding an  $L_1$  penalty to the LTS objective function, similar as for the lasso regression estimator (Tibshirani, 1996). The sparse LTS estimator can be applied to high-dimensional data, where the sample size exceeds the number of variables and is therefore as well a regularized version of LTS.

## 3 Robust Sparse alternating regression algorithm

We use a sequential algorithm to derive the canonical vectors.

*First canonical vector pair.* Denote the first canonical vector pair by  $(\mathbf{A}_1, \mathbf{B}_1)$ . Assume that the value of  $\mathbf{A}_1$  is known. Denote the vector of squared residuals by  $\mathbf{r}^2(\mathbf{B}_1) = (r_1^2, \dots, r_n^2)^T$ , with  $r_i^2 = (\mathbf{A}_1^T \mathbf{x}_i - \mathbf{B}_1^T \mathbf{y}_i)^2$ ,  $i = 1, \dots, n$ . The estimate of  $\mathbf{B}_1$  is obtained as

$$\hat{\mathbf{B}}_1 | \mathbf{A}_1 = \underset{\mathbf{B}_1}{\operatorname{argmin}} \sum_{i=1}^h (\mathbf{r}^2(\mathbf{B}_1))_{i:n} + \lambda_{B_1} \sum_{j=1}^p |\mathbf{B}_{1j}|, \quad (2)$$

where  $\lambda_{B_1} > 0$  is a sparsity parameter,  $\mathbf{B}_{1j}$  is the  $j^{th}$  element,  $j = 1, \dots, p$ , of the first canonical vector  $\mathbf{B}_1$ , and  $(\mathbf{r}^2(\mathbf{B}_1))_{1:n} \leq \dots \leq (\mathbf{r}^2(\mathbf{B}_1))_{n:n}$  are the order statistics of the squared residuals. The canonical vector  $\hat{\mathbf{B}}_1$  is normed to length 1. The solution to (2) equals the sparse LTS estimator with  $\mathbf{X}\mathbf{A}_1$  as response and  $\mathbf{Y}$  as predictor. Regularization by adding a penalty term to the objective function is necessary since the design matrix  $\mathbf{Y}$  can be high-dimensional. As such, we get a robust sparse estimate  $\hat{\mathbf{B}}_1$ .

Analogously, for a fixed value  $\mathbf{B}_1$ , denote the vector of squared residuals by  $\mathbf{r}^2(\mathbf{A}_1) = (r_1^2, \dots, r_n^2)^T$ , with  $r_i^2 = (\mathbf{B}_1^T \mathbf{y}_i - \mathbf{A}_1^T \mathbf{x}_i)^2, i = 1, \dots, n$ . The sparse LTS regression estimate of  $\mathbf{A}_1$  with  $\mathbf{Y}\mathbf{B}_1$  as response and  $\mathbf{X}$  as predictor, is

$$\hat{\mathbf{A}}_1 | \mathbf{B}_1 = \underset{\mathbf{A}_1}{\operatorname{argmin}} \sum_{i=1}^h (\mathbf{r}^2(\mathbf{A}_1))_{i:n} + \lambda_{A_1} \sum_{j=1}^q |\mathbf{A}_{1j}|, \quad (3)$$

where  $\lambda_{A_1} > 0$  is a sparsity parameter,  $\mathbf{A}_{1j}$  is the  $j^{th}$  element,  $j = 1, \dots, q$  of the first canonical vector  $\mathbf{A}_1$ , and  $(\mathbf{r}^2(\mathbf{A}_1))_{1:n} \leq \dots \leq (\mathbf{r}^2(\mathbf{A}_1))_{n:n}$  are the order statistics of the squared residuals. The canonical vector  $\hat{\mathbf{A}}_1$  is normed to length 1.

This leads to an alternating regression scheme, updating in each step the estimates of the canonical vectors until convergence. We iterate until the angle between the estimated canonical vectors in two successive iterations is smaller than some tolerance value  $\epsilon$  (e.g.  $\epsilon = 10^{-3}$ ), and this for  $\hat{\mathbf{A}}_1$  and  $\hat{\mathbf{B}}_1$ . In the simulations we conducted, convergence was almost always reached.

*Higher order canonical vector pairs.* We use deflated data matrices to estimate the higher order canonical vector pairs (see e.g. Branco et al., 2005). For the second canonical vector pair, the deflated matrices are  $\mathbf{X}^*$ , the residuals of a column-by-column LTS regression of  $\mathbf{X}$  on all lower order canonical variates,  $\hat{\mathbf{u}}_1$  in this case; and  $\mathbf{Y}^*$ , the residuals of a column-by-column LTS regression of  $\mathbf{Y}$  on  $\hat{\mathbf{v}}_1$ . Since these regressions only involve a small number of regressors, they should not be regularized.

The second canonical variate pair is then obtained by alternating between the following regressions until convergence:

$$\hat{\mathbf{B}}_2^* | \mathbf{A}_2^* = \underset{\mathbf{B}_2^*}{\operatorname{argmin}} \sum_{i=1}^h (\mathbf{r}^2(\mathbf{B}_2^*))_{i:n} + \lambda_{B_2^*} \sum_{j=1}^p |\mathbf{B}_{2j}^*|, \quad (4)$$

where  $\mathbf{r}^2(\mathbf{B}_2^*) = (r_1^2, \dots, r_n^2)^T$ , with  $r_i^2 = (\mathbf{A}_2^{*T} \mathbf{x}_i^* - \mathbf{B}_2^{*T} \mathbf{y}_i^*)^2, i = 1, \dots, n$ .

$$\hat{\mathbf{A}}_2^* | \mathbf{B}_2^* = \underset{\mathbf{A}_2^*}{\operatorname{argmin}} \sum_{i=1}^h (\mathbf{r}^2(\mathbf{A}_2^*))_{i:n} + \lambda_{A_2^*} \sum_{j=1}^q |\mathbf{A}_{2j}^*|, \quad (5)$$

where  $\mathbf{r}^2(\mathbf{A}_2^*) = (r_1^2, \dots, r_n^2)^T$ , with  $r_i^2 = (\mathbf{B}_2^{*T} \mathbf{y}_i^* - \mathbf{A}_2^{*T} \mathbf{x}_i^*)^2, i = 1, \dots, n$ . The canonical vectors  $\hat{\mathbf{B}}_2^*$  and  $\hat{\mathbf{A}}_2^*$  are both normed to length 1.

Finally, the second canonical vector needs to be expressed as linear combinations of the columns of the original data matrices, and not the deflated ones. Since we want to allow for zero coefficients in these linear combinations, a sparse approach is needed. To obtain a sparse  $\hat{\mathbf{A}}_2$ , we regress  $\hat{\mathbf{u}}_2^*$  on  $\mathbf{X}$  using the sparse LTS

estimator, yielding the fitted values  $\hat{\mathbf{u}}_2 = \mathbf{X}\hat{\mathbf{A}}_2$ . To obtain a sparse  $\hat{\mathbf{B}}_2$ , we regress  $\hat{\mathbf{v}}_2^*$  on  $\mathbf{Y}$  using the sparse LTS estimator, yielding the fitted values  $\hat{\mathbf{v}}_2 = \mathbf{Y}\hat{\mathbf{B}}_2$ .

The higher order canonical variate pairs are obtained in a similar way. We perform alternating sparse LTS regressions as in (4) and (5), followed by a final sparse LTS step to retrieve the estimated canonical vectors  $(\hat{\mathbf{A}}_{\mathbf{k}}, \hat{\mathbf{B}}_{\mathbf{k}})$ . Note that the regressions in (4) and (5) should be regularized, since the number of regressors equals  $p$  or  $q$ , which could be larger than  $n$ , but sparsity is not really necessary.

*Initial value.* A starting value for  $\mathbf{A}_1$  is required to start up the algorithm. Following Branco et al. (2005), we first obtain the first principal component of  $\mathbf{Y}$ , denoted  $\mathbf{z}_1$ . For this aim, we use the algorithm of Filzmoser et al. (2009) (available in the R package `pcaPP`) which performs robust principal component analysis. We regress  $\mathbf{z}_1$  on  $\mathbf{X}$  using the LTS estimator, or the sparse LTS if the number of variables is larger than the sample size. The estimated regression coefficient matrix of this regression is used as initial value for  $\mathbf{A}_1$ .

*Number of canonical variates to extract.* To decide on the number of canonical variates  $r$  to extract, we use the maximum eigenvalue ratio criterion of An et al. (2013). We apply the Robust Sparse CCA algorithm and calculate the robust correlations  $\hat{\rho}_1, \dots, \hat{\rho}_{\text{rmax}}$ , with  $\text{rmax} = \min(p, q)$ . Each  $\hat{\rho}_j$  is obtained by computing the correlation between  $\hat{\mathbf{v}}_j$  and  $\hat{\mathbf{u}}_j$  from the bivariate Minimum Covariance Determinant estimator with 25% trimming. Let  $\hat{k}_j = \hat{\rho}_j / \hat{\rho}_{j+1}$  for  $j = 1, \dots, \text{rmax} - 1$ . We extract  $r$  pairs of canonical variates, where  $r = \arg\max_j \hat{k}_j$ .

*Reweighted sparse LTS.* The sparse LTS estimator is computed with trimming proportion 25%, so size of the subsample  $h = \lfloor 0.75n \rfloor$ . To increase efficiency, we use a reweighting step afterwards. The reweighted sparse LTS is the lasso estimator computed from the observations not detected as outliers by the sparse LTS, i.e. having an absolute value of the standardized residuals smaller than or equal to the 98.75<sup>th</sup> quantile of the standard normal distribution (see Alfons et al., 2013 for more detail).

*Choice of the sparsity parameter.* The sparsity parameter controlling the penalization on the regression coefficient matrices is selected with the Bayesian Information Criterion (e.g. Yin and Li, 2011). We use a range of values for the sparsity parameter  $\lambda$  and select the one with the lowest value of

$$BIC_\lambda = -2 \log L_\lambda + k_\lambda \log(n),$$

where  $L_\lambda$  is the estimated likelihood, using sparsity parameter  $\lambda$  and  $k_\lambda$  is the number of non-zero estimated regression coefficients.

Table 1: Simulation settings. In all settings  $\Sigma_{xx} = \mathbf{I}_p$  and  $\Sigma_{yy} = \mathbf{I}_q$ .

Simulation setting	$n$	$p$	$q$	$\Sigma_{xy}$
Sparse Low-dimensional	100	6	4	$\begin{bmatrix} 0.9 & \mathbf{0}_{1 \times 3} \\ \mathbf{0}_{5 \times 1} & \mathbf{0}_{5 \times 3} \end{bmatrix}$
Non-Sparse Low-dimensional	250	12	8	$\begin{bmatrix} 0.2 & \mathbf{0.1}_{1 \times 7} \\ \mathbf{0.1}_{11 \times 1} & \mathbf{0.1}_{11 \times 7} \end{bmatrix}$
Sparse High-dimensional	100	100	4	$\begin{bmatrix} \mathbf{0.45}_{2 \times 2} & \mathbf{0}_{2 \times 2} \\ \mathbf{0}_{98 \times 2} & \mathbf{0}_{98 \times 98} \end{bmatrix}$

## 4 Simulation Study

We compare the performance of the Robust Sparse CCA method with (i) standard CCA, (ii) Robust CCA, and (iii) Sparse CCA. The same algorithm is used for all 4 estimators, for ease of comparability. Robust CCA uses LTS instead of sparse LTS, and corresponds to the alternating regression approach of Branco et al. (2005). Standard CCA uses LS instead of LTS, Pearson correlations for computing the canonical correlations, and ordinary PCA for getting the initial values. Sparse CCA is like standard CCA, but used the lasso instead of LS for the multiple regressions.

Several simulation settings are considered. In the first simulation design (revised from Branco et al., 2005), there is one canonical variate pair and the canonical vectors have a sparse structure. The canonical vectors are very sparse; each containing only one non-zero element. In the second design, there are two canonical variate pairs and the canonical vectors are non-sparse. In the third design, there are a lot of variables ( $p = 100$ ) compared to the sample size ( $n = 100$ ). There is one canonical variate pair and the canonical vectors are sparse. Only Sparse CCA and Robust Sparse CCA can be computed. The number of simulations for each setting is  $M = 500$ .

For each setting, the following sampling distributions are considered (Branco et al., 2005)

- (a) *No contamination.* We generate data matrices  $\mathbf{X}$  and  $\mathbf{Y}$  according to multivariate normal distributions  $N_{p+q}(\mathbf{0}, \mathbf{\Sigma})$ , with covariance matrices described in Table 1.
- (b) *Symmetric contamination.* 90% of the data are generated from  $N_{p+q}(\mathbf{0}, \mathbf{\Sigma})$ , and 10% of the data are generated from  $N_{p+q}(\mathbf{0}, 9\mathbf{\Sigma})$ .
- (c) *Asymmetric contamination.* 90% of the data are generated from  $N_{p+q}(\mathbf{0}, \mathbf{\Sigma})$ , and 10% of the observations equal the point  $\text{tr}(\mathbf{\Sigma})\mathbf{1}^T$ , where  $\text{tr}(\mathbf{\Sigma})$  is the trace of  $\mathbf{\Sigma}$ .

**4.1 Performance measures.** The estimators are evaluated on their estimation accuracy. We compute for each simulation run  $m$ , with  $m = 1, \dots, M = 500$ , the angle  $\theta^m(\hat{\mathbf{A}}^m, \mathbf{A})$  between the subspace spanned by the estimated canonical vectors (contained in the columns of  $\hat{\mathbf{A}}^m$ ) and the subspace spanned by the true canonical vectors (contained in the columns of  $\mathbf{A}$ ).<sup>1</sup> Analogously for the matrix  $\mathbf{B}$ . The median angles are given by

$$\theta_{\text{Med}}(\hat{\mathbf{A}}, \mathbf{A}) = \text{med}_{1 \leq m \leq M} \theta^m(\hat{\mathbf{A}}^m, \mathbf{A}) \quad \text{and} \quad \theta_{\text{Med}}(\hat{\mathbf{B}}, \mathbf{B}) = \text{med}_{1 \leq m \leq M} \theta^m(\hat{\mathbf{B}}^m, \mathbf{B}).$$

For evaluating sparsity, we use the true positive rate and the true negative rate (e.g. Rothman et al., 2010)

$$\begin{aligned} \text{TPR}(\hat{\mathbf{A}}^m, \mathbf{A}) &= \frac{\#\{(i, j) : \widehat{\mathbf{A}}_{ij}^m \neq 0 \text{ and } \mathbf{A}_{ij} \neq 0\}}{\#\{(i, j) : \mathbf{A}_{ij} \neq 0\}} \\ \text{TNR}(\hat{\mathbf{A}}^m, \mathbf{A}) &= \frac{\#\{(i, j) : \widehat{\mathbf{A}}_{ij}^m = 0 \text{ and } \mathbf{A}_{ij} = 0\}}{\#\{(i, j) : \mathbf{A}_{ij} = 0\}}. \end{aligned}$$

Analogously for the matrix  $\mathbf{B}$ . A true positive is a coefficient that is non-zero in the true model, and is estimated as non-zero. A true negative is a coefficient that is zero in the true model, and is estimated as zero. Both the true positive rate and the true negative rate should be as high as possible for a sparse estimator.

**4.2 Results for the Sparse Low-dimensional design.** Simulation results for the estimator  $\hat{\mathbf{A}}$  are in Table 2. The results for the estimator  $\hat{\mathbf{B}}$  are similar and are, therefore, omitted. In the scenario without contamination, the results are according to the expectations. The sparse estimators Sparse CCA and Robust Sparse CCA achieve a much better median estimation accuracy than the non-sparse estimators CCA and Robust CCA. As expected, a sparse method results in increased estimation accuracy when the true canonical vectors have a sparse structure. Looking at sparsity recognition performance, Sparse CCA and Robust Sparse CCA perform equally good in retrieving the sparsity in the data generating process.

In the contaminated simulation settings, we see that the robust estimators Robust Sparse CCA and Robust CCA maintain their accuracy under contamination. Robust Sparse CCA performs best. Robust

---

<sup>1</sup>The (minimum) angle  $\theta(\hat{\mathbf{A}}, \mathbf{A})$  is computed as follows. Compute the QR-decompositions  $\hat{\mathbf{A}} = \mathbf{Q}_{\hat{\mathbf{A}}} \mathbf{R}_{\hat{\mathbf{A}}}$  and  $\mathbf{A} = \mathbf{Q}_{\mathbf{A}} \mathbf{R}_{\mathbf{A}}$ . Next, compute the singular value decomposition of  $\mathbf{Q}_{\hat{\mathbf{A}}}^T \mathbf{Q}_{\mathbf{A}} = \mathbf{U} \mathbf{C} \mathbf{V}^T$ . The matrix  $\mathbf{C}$  is diagonal with elements  $c_1, \dots, c_r$ . The minimum angle is given by  $\theta(\hat{\mathbf{A}}, \mathbf{A}) = \arccos(c_1)$ .



Table 2: Results for the Sparse Low-dimensional design. Median of the angles between the space spanned by the true and estimated canonical vectors; median true positive rate and true negative rate are reported for each method. Averages of the angles are between parentheses.

Method	No contamination			Symmetric contamination			Asymmetric contamination		
	$\theta_{\text{Med}}(\hat{\mathbf{A}}, \mathbf{A})$	TPR	TNR	$\theta_{\text{Med}}(\hat{\mathbf{A}}, \mathbf{A})$	TPR	TNR	$\theta_{\text{Med}}(\hat{\mathbf{A}}, \mathbf{A})$	TPR	TNR
CCA	0.08 (0.08)			0.16 (0.18)			0.24 (0.24)		
Robust CCA	0.10 (0.11)			0.14 (0.14)			0.14 (0.14)		
Sparse CCA	0.00 (0.02)	1.00 (0.98)	1.00 (0.97)	0.15 (0.15)	1.00 (1.00)	0.20 (0.19)	0.19 (0.19)	1.00 (1.00)	0.01 (0.10)
Robust Sparse CCA	0.00 (0.03)	1.00 (0.99)	1.00 (0.89)	0.01 (0.03)	1.00 (0.99)	0.80 (0.84)	0.00 (0.05)	1.00 (0.98)	1.00 (0.88)

Table 3: Results for the Non-sparse Low-dimensional design. Median of the angles between the space spanned by the true and estimated canonical vectors are reported for each method. Averages of the angles are between parentheses.

Method	No contamination	Symmetric contamination	Asymmetric contamination
	$\theta_{\text{Med}}(\hat{\mathbf{A}}, \mathbf{A})$	$\theta_{\text{Med}}(\hat{\mathbf{A}}, \mathbf{A})$	$\theta_{\text{Med}}(\hat{\mathbf{A}}, \mathbf{A})$
CCA	0.03 (0.03)	0.09 (0.18)	0.10 (0.33)
Robust CCA	0.04 (0.04)	0.04 (0.04)	0.04 (0.05)
Sparse CCA	0.03 (0.03)	0.11 (0.21)	0.09 (0.08)
Robust Sparse CCA	0.04 (0.04)	0.04 (0.04)	0.05 (0.05)

Sparse CCA clearly outperforms Robust CCA: for instance, we have a median estimation accuracy of 0.01 (0.00) against 0.14 (0.14) for symmetric (asymmetric) contamination, see Table 2. The non-robust estimators CCA and Sparse CCA are clearly influenced by the introduction of outliers, as reflected by the much higher values of the median angles  $\theta_{\text{Med}}(\hat{\mathbf{A}}, \mathbf{A})$ . Sparse CCA now performs even worse than Robust CCA. The non-robust estimators are most influenced in the more severe asymmetric contamination setting. Looking at sparsity recognition performance, Robust Sparse CCA method is hardly affected by the introduction of outliers. For the Sparse CCA method, especially the true negative rates is affected.

**4.3 Results for the Non-sparse Low-dimensional design.** Summarizing results are in Table 3. Note that the true positive rate and true negative rate are omitted since the true canonical vectors are non-sparse. In the situation without contamination, all methods show similar performance. The price the sparse methods pay is a marginally decreased estimation accuracy, as measured by the median and average angle. In the contaminated settings, the robust methods Robust Sparse CCA and Robust CCA perform best and show similar performance.

Table 4: Results for Sparse High-dimensional design. Median of the angles between the space spanned by the true and estimated canonical vectors; median true positive rate and true negative rate are reported for each method. Averages of the angles are between parentheses.

Method	No contamination			Symmetric contamination			Asymmetric contamination		
	$\theta_{\text{Med}}(\hat{\mathbf{A}}, \mathbf{A})$	TPR	TNR	$\theta_{\text{Med}}(\hat{\mathbf{A}}, \mathbf{A})$	TPR	TNR	$\theta_{\text{Med}}(\hat{\mathbf{A}}, \mathbf{A})$	TPR	TNR
Sparse CCA	0.06 (0.13)	1.00 (0.96)	0.98 (0.97)	0.19 (0.37)	1.00 (0.92)	0.88 (0.86)	0.33 (0.34)	1.00 (1.00)	1.00 (0.98)
Robust Sparse CCA	0.08 (0.35)	1.00 (0.83)	0.99 (0.98)	0.11 (0.36)	1.00 (0.84)	0.98 (0.98)	0.08 (0.33)	1.00 (0.82)	1.00 (0.99)

**4.4 Results for Sparse High-dimensional design.** Summarizing results are in Table 4. In this high-dimensional setting, only Sparse CCA and Robust Sparse CCA can be performed. In the scenario without contamination, Sparse CCA performs best. Sparse CCA is, however, closely followed by Robust Sparse CCA both in terms of median estimation accuracy and sparsity recognition performance. When adding contamination, the performance of Sparse CCA gets distorted. The median estimation accuracy of Robust Sparse CCA is much better than the median estimation accuracy of Sparse CCA, especially in the more severe asymmetric contamination setting.

Similar conclusions can be made when looking at the average angle as a measure of estimation accuracy (reported in Tables 2 to 4 between parentheses). In sum, Robust Sparse CCA shows the best overall performance in this simulation study. Robust Sparse CCA combines both robustness and sparsity. Therefore, it performs best in sparse settings where contamination is present. In sparse non-contaminated settings, Robust Sparse CCA shows competitive performance to its main competitor Sparse CCA. In contaminated non-sparse settings, Robust Sparse CCA shows competitive performance to its main competitor Robust CCA.

## 5 Applications

**5.1 Evaporation data set.** As a first example, we use a data set from Freund (1979). Ten variables (maximum, minimum and average soil temperature; maximum, minimum and average air temperature; maximum, minimum and average daily relative humidity; and total wind) have been measured on 46 consecutive days from June 6 until July 21. The aim is to find and quantify the relations between the soil temperature variables and the remaining variables.

As a first inspection of the data, we use the Distance-Distance plot (Rousseeuw and van Zomeren, 1990) in Figure 1. The Distance-distance plot displays the robust distances versus the Mahalanobis distances. The vertical and horizontal lines are drawn at values equal to the square root of the 97.5% quantile of a chi-

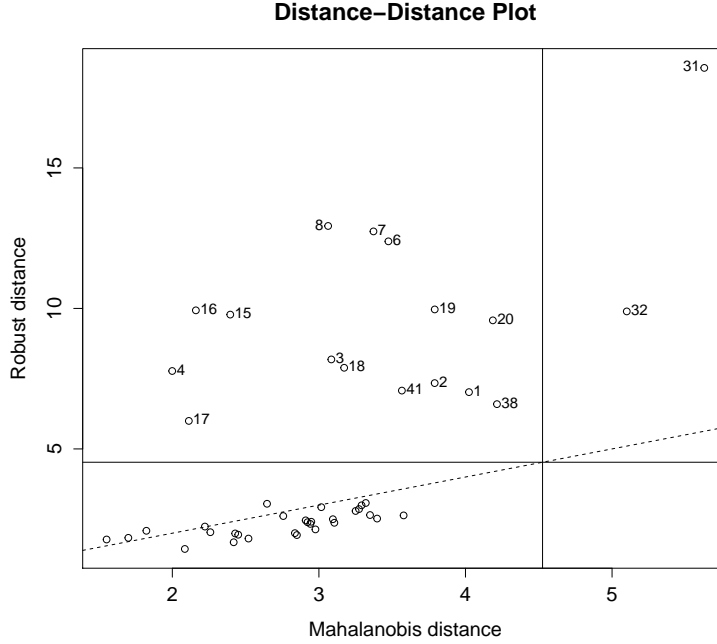


Figure 1: Distance-Distance Plot for the Evaporation data

squared distribution with 10 degrees of freedom. Points beyond those lines would be considered as outliers. The Distance-Distance plot reveals some outliers: objects 31 and 32, for example, are extreme outliers. This suggests the need for a robust CCA method.

We use the maximum eigenvalue ration criterion (e.g. An et al., 2013), as discussed in Section 3, to decide on the number of canonical variate pairs to extract. For all CCA methods two canonical variate pairs are extracted. To compare the performance of the CCA approaches, we perform a leave-one-out cross-validation exercise and compute the cross-validation score

$$CV = \frac{1}{h} \sum_{i=1}^h \|\hat{\mathbf{A}}_{-i}^T \mathbf{x}_i - \hat{\mathbf{B}}_{-i}^T \mathbf{y}_i\|^2,$$

where  $\hat{\mathbf{A}}_{-i}^T$  and  $\hat{\mathbf{B}}_{-i}^T$  contain the estimated canonical vectors when the  $i^{th}$  observation is left out of the estimation sample and  $h = \lfloor n\alpha \rfloor$ , with  $\alpha = 1$  (0% Trimming) or  $\alpha = 0.9$  (10% Trimming). We use trimming to eliminate the effect of outliers in the cross-validation score. The method that achieves the lowest cross-validation score has the best out-of-sample performance. Table 5 reports the cross-validation scores for all methods. Robust Sparse CCA achieves the best cross-validation score.

Table 6 shows the estimated canonical vectors for the Robust CCA and Robust Sparse CCA method. By adding the penatly term, the number of non-zero coefficients is reduced from 20 to 10, improving interpretability of the canonical vectors. The price to pay for the sparseness is a slight decrease in the estimated

Table 5: Evaporation data set: Cross-validation score for each method.

Method	CV-score	
	0% Trimming	10% Trimming
CCA	1.48	0.98
Robust CCA	1.26	0.73
Sparse CCA	1.63	0.68
Robust Sparse CCA	0.90	0.66

Table 6: Evaporation data set: Estimated canonical vectors using Robust CCA and Robust Sparse CCA.

Variables \ Canonical Vectors	Robust CCA		Robust Sparse CCA	
	1	2	1	2
<i>First data set</i>				
MAXST: Max. daily soil temperature	-0.21	-0.73	0	0.76
MINST: Min. daily soil temperature	-0.02	0.67	0	-0.63
AVST: Avg. daily soil temperature	0.98	0.13	1	-0.15
<i>Second data set</i>				
MAXAT: Max. daily air temperature	0.61	0.01	0.94	0
MINAT: Min. daily air temperature	0.53	0.75	0.15	-0.84
AVAT: Avg. daily air temperature	0.08	-0.05	0.17	0
MAXH: Max. daily relative humidity	-0.11	0.05	0	0
MINH: Min. daily relative humidity	0.16	0.54	0	-0.53
AVH: Avg. daily relative humidity	-0.43	0.33	-0.24	0
WIND: Total wind, measured in miles per day	-0.35	0	0	0
<i>Canonical correlations</i>	0.89	0.59	0.87	0.48

Table 7: Nutrimouse data set: Cross-validation score for each method.

Method	CV-score	CV-score
	0% Trimming	10% Trimming
Sparse CCA	190.89	114.42
Robust Sparse CCA	78.57	37.10

canonical correlations (computed using the bivariate MCD estimator, see Section 3): they drop from 0.89 to 0.87 for the first one, and from 0.59 to 0.48 for the second canonical correlation. We find this decrease acceptable, given the gained sparsity in the canonical vectors. The sparse structure of the canonical vectors facilitates interpretation. The first canonical variate in the soil temperature data set, for instance, is uniquely determined by the variable AVST.

**5.2 Nutrimouse data set.** As a second example, we use the nutrimouse data set (publicly available in the R package *CCA*; Gonzalez et al., 2008). Two sets of variables, i.e. gene expressions and fatty acids, are available for 40 mice. The first set contains expressions of 120 genes measured in liver cells. The second set of variables contains concentrations of 21 hepatic fatty acids (FA). Furthermore, the mice are classified according to two factors: genotypes (wild-type and PPAR $\alpha$  deficient mice) and diets (reference diet “REF”, hydrogenated coconut oil diet “COC”, sunflower oil diet “SUN”, linseed oil diet “LIN” and fish oil diet “FISH”). More details on how the data were obtained can be found in Martin et al. (2007). The aim is to identify a small set of genes which are correlated with the fatty acids.

In this data set, the number of experimental units (i.e. data on 40 mice) is smaller than the number of variables (i.e. 120 genes). Therefore, standard CCA nor robust CCA can be performed. Gonzalez et al. (2008) use the Canonical Ridge. The Canonical Ridge performs regularization, but the canonical vectors produced by the Canonical Ridge are not sparse. Robust Sparse CCA and Sparse CCA, can also be applied in this high-dimensional setting and produce - more interpretable - sparse canonical vectors.

We compare the performance of the Robust Sparse CCA method to the Sparse CCA method. For both methods, two canonical variate pairs are extracted using the maximum eigenvalue ration criterion from Section 3. We perform the same leave-one-out cross-validation exercise as described in Section 5.1. Results are reported in Table 7. The Robust Sparse CCA method outperforms the Sparse CCA method. The cross-validation scores are reduced by about 60% when using the robust method.

Next, we discuss the estimated canonical vectors obtained using the Robust Sparse CCA method. The most informative biological conclusions can be drawn from (i) the selected genes in the first canonical vector and (ii) the selected fatty acids in the second canonical vector (for further motivation see, Martin et al., 2007; Gonzalez et al., 2008). Therefore, we focus on these results. First, the left panel of Figure 2 displays the

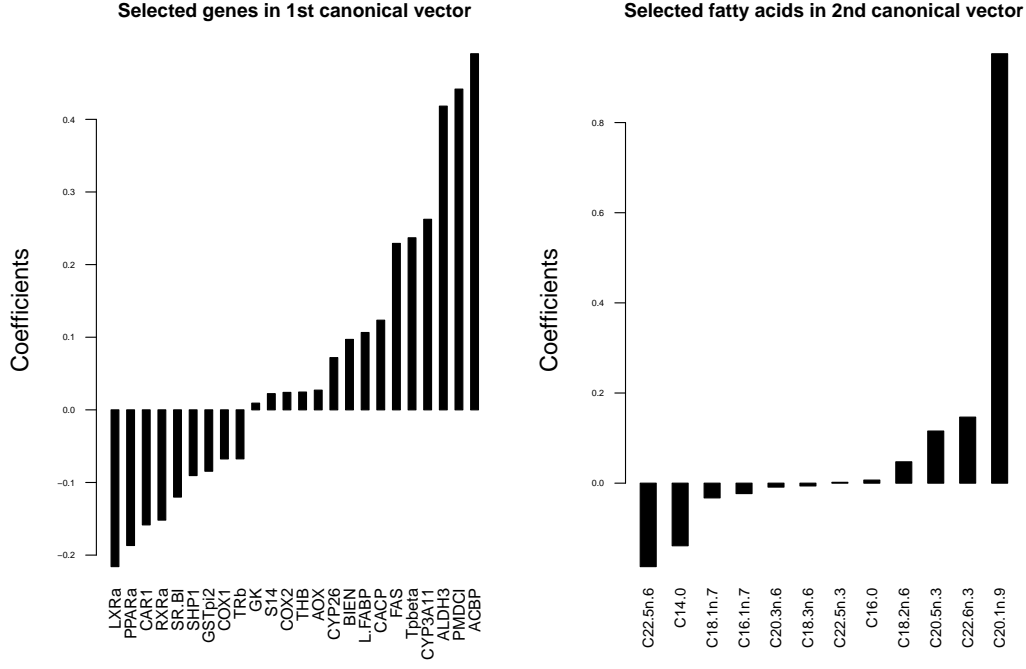


Figure 2: Coefficients of selected genes (left) and coefficients of selected fatty acids (right).

coefficients of the selected genes, i.e. those genes with non-zero estimated coefficients, in the first canonical vector. 24 out of 120 variables are selected. The solution is very sparse, facilitating interpretation. Two important genes are *Cyp3a11* and *CAR1*. Martin et al. (2007) find a consistent reduction of *Cyp3a11* in *PPARα* livers on the one hand, and an overexpression of *CAR1* on the other hand. Both genes are selected and have among the highest (absolute) coefficients. Second, we focus on the fatty acids and discuss the right panel of Figure 2. The right panel displays the coefficients of the selected fatty acids in the second canonical vector. 12 out of 21 fatty acid variables are selected. The fatty acids are related to the effect of certain diets used in this experiment. The mice are classified according to a specific diet. Five diets are used, which contain specific concentration levels of the fatty acids *C22:6n-3*, *C22:5n-3*, *C22:5n-6*, *C22:4n-3* and *C20:5n-3* respectively (Martin et al., 2007). Looking at the selected fatty acids in the second canonical vector, we see that four out of these five variables are selected and have among the highest (absolute) coefficients.

In sum, the strong genotype effect observed through the first canonical variate and the strong diet effect observed through the second canonical variate is in accordance with conclusions drawn in Martin et al. (2007).

## 6 Conclusion

Sparse Canonical Correlation Analysis delivers interpretable canonical vectors, with some of its elements estimated as exactly zero. Robust Sparse CCA retains this advantage, while at the same ensuring that the estimation of the canonical vectors is not affected by outlying observations.

The canonical vectors are given by the eigenvectors of two particular matrices, see for instance Johnson and Wichern (1998, Chapter 10). Typically, the canonical vectors are estimated by taking the sample versions of those covariance matrices and computing the corresponding eigenvectors. One could think of estimating those covariance matrices with an estimator that is robust and sparse at the same time, and then, to compute the eigenvectors. This approach, however, would results in canonical vectors being non-sparse. To circumvent this pitfall, we reformulate the CCA problem in a regression framework. We use the sparse LTS estimator of Alfons et al. (2013) to obtain sparse canonical vectors that are resistant to outlying observations.

A simulation study and two empirical examples show the advantages of the Robust Sparse CCA method over its natural benchmarks. In contaminated settings, Robust Sparse CCA outperforms standard CCA and Sparse CCA. Robust Sparse CCA has two important advantages over Robust CCA. First, Robust Sparse CCA improves model interpretation since only a limited number of variables, those corresponding to the non-zero elements of the canonical vectors, enter the estimated canonical variates (cfr. evaporation application). Second, if the number of variables approaches the sample size, the estimation precision of Robust CCA suffers. If the number of variables exceeds the sample size, Robust CCA can not even be performed. Due to the regularization performed by Robust Sparse CCA, Robust Sparse CCA can still be computed (cfr. nutrimouse application).

Several questions are left for future research. One could think of a joint selection criterion for the number of canonical variate pairs and the sparsity parameter. This would, however, increase computation time substantially. To induce sparsity in the canonical vectors, we use a Lasso penalty. Other penalty functions such as the Adaptive Lasso (Zou, 2006) could be considered. The Adaptive Lasso is consistent for variable selection, whereas the Lasso is not. Furthermore, we use a regularized version of the LTS estimator. One could also use a regularized version of the S-estimator or the MM-estimator to increase efficiency. Up to our knowledge, however, the sparse LTS is the only robust sparse regression estimator for which efficient code (Alfons, 2014) is available.

## Acknowledgments

Financial support from the FWO (Research Foundation Flanders) is gratefully acknowledged (FWO, contract number 11N9913N).

## References

- Adrover, J. and Donato, S. (2015), “A robust predictive approach for canonical correlation analysis,” *Journal of Multivariate Analysis*, 133, 356–376.
- Alfons, A. (2014), *robustHD: Robust methods for high-dimensional data*, R package version 0.5.0.
- Alfons, A.; Croux, C. and Gelper, S. (2013), “Sparse least trimmed squares regression for analyzing high-dimensional large data sets,” *The Annals of Applied Statistics*, 7, 226–248.
- An, B.; Guo, J. and Wang, H. (2013), “Multivariate regression shrinkage and selection by canonical correlation analysis,” *Computational Statistics and Data Analysis*, 62, 93–107.
- Branco, J.; Croux, C.; Filzmoser, P. and Oliviera, M. (2005), “Robust Canonical Correlations: A Comparative Study,” *Computational Statistics*, 20, 203–229.
- Brillinger, D. (1975), *Time Series: Data analysis and theory*, New York: Holt, Rinehart, and Winston.
- Cruz-Cano, R. and Lee, M.-L. (2014), “Fast regularized canonical correlation analysis,” *Computational Statistics and Data Analysis*, 70, 88–100.
- Dehon, C. and Croux, C. (2002), “Analyse canonique basée sur des estimateurs robustes de la matrice de covariance,” *La Revue de Statistique Appliquée*, 2, 5–26.
- Filzmoser, P.; Croux, C. and Dehon, C. (2000), “Outlier Resistant Estimators for Canonical Correlation Analysis,” *Compstat: Proceedings in Computational Statistics*, eds. J.G. Bethlehem and P.G.M. van de Heijden, Heidelberg, Physica Verlag, 301–307.
- Filzmoser, P.; Fritz, H. and Kalcher, K. (2009), *pcaPP: Robust PCA by Projection Pursuit*, R package version 1.9.
- Freund, R. (1979), “Multicollinearity etc. Some ‘new’ examples,” *American Statistical Association Proceedings of Statistical Computing Section*, 111–112.
- Gonzalez, I.; Dejean, S.; Martin, P. and Baccini, A. (2008), “CCA: An R package to extend canonical correlation analysis,” *Journal of Statistical Software*, 23, 1–14.
- Hotelling, H. (1936), “Relations between two sets of variates,” *Biometrika*, 28, 321–377.
- Izenman, A. (1975), “Reduced-rank regression for the multivariate linear model,” *Journal of Multivariate Analysis*, 5, 248–264.
- Johnson, R. and Wichern, D. (1998), *Applied Multivariate Statistical Analysis*, London: Prentice-Hall.
- Karnel, G. (1991), *Robust canonical correlation and correspondence analysis*, In: The Frontiers of Statistical Scientific and Industrial Applications. Volume II of the proceedings of ICOSCO-I, The First International Conference on Statistical Computing, American Sciences Press, Strassbourg.
- Kudraszow, N. and Maronna, R. (2011), “Robust canonical correlation analysis: a predictive approach,” *Working paper*.



- Lykou, A. and Whittaker, J. (2010), “Sparse CCA using a lasso with positivity constraints,” *Computational Statistics and Data Analysis*, 54, 3144–3157.
- Martin, P.; Guillon, H.; Lasserre, F.; Dejean, S.; Lan, A.; Pascussi, J.; SanCristobal, M.; Legrand, P.; Besse, P. and Pineau, T. (2007), “Novel aspects of PPAR $\alpha$ -mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study,” *Hepatology*, 54, 767–777.
- Parkhomenko, E.; Tritchler, D. and Beyene, J. (2009), “Sparse canonical correlation analysis with application to genomic data integration,” *Statistical Applications in Genetics and Molecular Biology*, 8, 1–34.
- Prabhakar, C. and Fridley, B. (2012), “Comparison of penalty functions for sparse canonical correlation analysis,” *Computational Statistics and Data Analysis*, 56, 245–254.
- Rothman, A.; Levina, E. and Zhu, J. (2010), “Sparse multivariate regression with covariance estimation,” *Journal of Computational and Graphical Statistics*, 19, 947–962.
- Rousseeuw, P. (1985), *Multivariate estimators with high breakdown point*, W. Grossman, G. Pflug, I. Vincza, W. Wertz (Eds.), Mathematical Statistics and its Applications, vol. B, Reidel, Dordrecht, The Netherlands.
- Rousseeuw, P. and van Zomeren, B. (1990), “Unmasking multivariate outliers and leverage points,” *Journal of the American Statistical Association*, 85, 633–639.
- Sun, L.; Ji, S. and Ye, J. (2011), “Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33, 194–200.
- Taskinen, S.; Croux, C.; Kankainen, A.; Ollila, E. and Oja, H. (2006), “Canonical Analysis based on Scatter Matrices,” *Journal of Multivariate Analysis*, 97, 359–384.
- Tibshirani, R. (1996), “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society Series B*, 58, 267–288.
- Waaaijenborg, S.; Hamer, P. and Zwinderman, A. (2008), “Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis,” *Statistical Applications in Genetics and Molecular Biology*, 7, Article 3.
- Wilms, I. and Croux, C. (2013), “Sparse canonical correlation analysis from a predictive point of view,” *FEB Research Report KBI 1320*, 1–24.
- Witten, D.; Tibshirani, R. and Hastie, T. (2009), “A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis,” *Biostatistics*, 10, 515–534.
- Wold, H. (1968), *Nonlinear estimation by iterative least square procedures*, New York: Wiley.
- Yin, J. and Li, H. (2011), “A sparse conditional gaussian graphical model for analysis of genetical genomics data,” *The Annals of Applied Statistics*, 5, 2630–2650.
- Zou, H. (2006), “The adaptive lasso and its oracle properties,” *Journal of the American Statistical Association*, 101, 1418–1429.

**FACULTY OF ECONOMICS AND BUSINESS**

Naamsestraat 69 bus 3500

3000 LEUVEN, BELGIË

tel. + 32 16 32 66 12

fax + 32 16 32 67 91

info@econ.kuleuven.be

www.econ.kuleuven.be

